Theoria, 2019, 34(3), 407-421
https://doi.org/10.1387/theoria.20040

THEORIA

ISSN 0495-4548 – eISSN 2171-679X

The philosopher's paradox: How to make a coherent decision in the Newcomb Problem

(La paradoja del filósofo: cómo hacer una decisión coherente en el problema de Newcomb)

Christopher VIGER[1], Carl HOEFER*[2,3], Daniel VIGER[4]

[1] Western University, Canada
[2] ICREA, Spain
[3] Universitat de Barcelona, Spain
[4] Western University, Canada

ABSTRACT: We offer a novel argument for one-boxing in Newcomb's Problem. The intentional states of a rational person are psychologically coherent across time, and rational decisions are made against this backdrop. We compare this coherence constraint with a golf swing, which to be effective must include a follow-through after the ball is in flight. Decisions, like golf swings, are extended processes, and their coherence with other psychological states of a player in the Newcomb scenario links her choice with the way she is predicted in a common cause structure. As a result, the standard argument for two-boxing is mistaken.

KEYWORDS: Newcomb's Problem; rationality; counterfactuals; psychological coherence; dominant strategy; common cause.

RESUMEN: *Ofrecemos un argumento novedoso a favor de elegir solo una caja en el Problema de Newcomb. Los estados intencionales de una persona racional son psicológicamente coherente a través del tiempo, y las decisiones están hechas con esta coherencia como trasfondo. Comparamos esta coherencia a través del tiempo con un* swing *de golf que, para ser efectivo, tiene que incluir un buen* follow through *cuando la bola ya está en el aire. Decisiones, como* swings *de golf, son procesos extendidos, y su coherencia con otros estados psicológicos del jugador en el escenario de Newcomb vincula su elección con la predicción hecha sobre ella en una estructura de causa común. En consecuencia, el argumento estándar para elegir dos cajas es equivocado.*

PALABRAS CLAVE: Problema de Newcomb; racionalidad; contrafácticos; coherencia psicológica; estrategia dominante; causa común.

**** Correspondence to:*** Carl Hoefer. ICREA, Pg. Lluís Companys 23 (08010 Barcelona), Spain. Departament de Filosofia, Universitat de Barcelona. Carrer Montalegre 6, 4.ª planta (08001 Barcelona), Spain – carl.hoefer@ub.edu – http://orcid.org/0000-0002-8020-4630

How to cite: Viger, Christopher; Hoefer, Carl; Viger; Daniel. (2019). «The philosopher's paradox: How to make a coherent decision in the Newcomb Problem»; *Theoria. An International Journal for Theory, History and Foundations of Science,* 34(3), 407-421. (https://doi.org/10.1387/theoria.20040).

Received: 6 July, 2018; Final version: 22 March, 2019.
ISSN 0495-4548 - eISSN 2171-679X / © 2019 UPV/EHU

1. *Introduction*

We offer a novel argument for choosing one box in Newcomb's Problem.[1] We begin with a brief vignette to present the essential features of Newcomb's Problem and to make the decision context more salient. We present a standard analysis of the problem due to Gibbard and Harper (1978) and make explicit some relevant assumptions about the predictor. Our analysis begins by observing that the intentional states of a rational person are essentially psychologically coherent and that rational decisions are made against this backdrop of personal ratio155nality. We elaborate this coherence constraint on rationality through an analogy with a golf swing, which to be effective includes a follow-through after the ball is in flight. Decisions, like golf swings, are extended processes, and their coherence with other psychological states of a player in the Newcomb scenario links her choice with the way she is predicted. As a result, the standard analysis according to which choosing two boxes is a dominant strategy is mistaken. In fact, that analysis would apply equally to someone who was never interviewed by a predictor at all, and simply found herself in front of two boxes. We argue that psychological coherence requires certain backtracking counterfactuals to be true in the scenario stipulated in the Newcomb Problem, which entail that choosing one box is the rational decision. In effect, a person's psychological make-up is a common cause of what she will choose and what is in the opaque box, mediated via the predictor. Treating decisions as isolated, independent events can, in certain circumstances such as the reflexive context of the Newcomb Problem, lead to paradoxes about what is rational.

2. *Newcomb's Problem (Vignette)*

> Host: Let's welcome our next two contestants on today's show. Our first contestant is a single working parent of three; the second is a graduate student in philosophy. Given your circumstances, clearly you both could really use as much money as you can possibly get today.
>
> Contestants 1 and 2 together: Yes, absolutely.
>
> Host: Okay, remember the rules of the game. We have two boxes, A and B. We can all see that box B (which is transparent) has $1,000 in it. Opaque Box A may or may not contain $1,000,000. You can choose either box A by itself or both boxes. And of course the catch is that before coming out on stage, you have each been interviewed by Cassandra, our oracle. Based on your interviews, she has predicted whether you will pick one or two boxes. If she predicted you will pick two boxes, then box A will be empty; but if she predicted you will pick just box A, it will contain the $1,000,000.
>
> <Crowd cheers wildly>
>
> Host: And remember Cassandra is correct in her predications 99% of the time. Contestant number 1, this is your big moment. So what are you going to choose?

[1] Other proponents of choosing one box include Bar-Hillel and Margalit (1972), Dummett (1993), Horgan (1981), Horowich (1985), Price (1986, 1991, 2012), Spohn (2012), and Vinci (1988). Ahmed (2014) and Hunter and Richter (1978) find fault with the causal decision theory (CDT) typically invoked to make the case for choosing two boxes. While we are sympathetic to many of the conclusions these authors reach, and argue for some ourselves as noted in the text, only Spohn's reasoning is directly relevant to the argument we present here.

> Contestant 1: Well I really need the money and most of the people who pick one box win $1,000,000 so that's my choice, just box A.
>
> During a dramatic pause, Contestant 2's classmates scoff at the irrationality of her decision. "Why would she leave $1,000 on the table?"[2] Just then box A is raised revealing $1,000,000 accompanied by flashing lights, sirens, and confetti. As the excitement subsides Contestant 1 leaves the stage with her $1,000,000.
>
> Host: Now Contestant 2, you're a graduate student in philosophy. Did Cassandra ask you about that?
>
> Contestant 2: Yes. That was the only question she asked me.
>
> <Host grimaces>
>
> Host: Well, you've seen how it's done. What's your choice?
>
> Contestant 2: Since $1,000,000 is already either in box A or it's not and I can't change that now, to get as much money as possible I choose both boxes. Why would I leave $1,000 on the table?
>
> Contestant 2's friends nod approvingly. Box A is raised revealing nothing. Contestant 2 is heard mumbling, "Good thing I took both boxes or I'd have gotten nothing" while exiting the stage with her $1,000.

3. *The Dominant Strategy*

Gibbard and Harper (1978) argue that the rational choice in Newcomb's Problem is to choose two boxes because at the time a decision is made what is in the opaque box has already been determined; the choice is not a cause of what is in the opaque box. Since the objective is to get as much money as possible, Gibbard and Harper reason that regardless of whether the opaque box contains one million dollars or nothing, the payout in each case is larger by taking both boxes, making that the rational choice; i.e. choosing two boxes is the dominant strategy.

> Rational choice in Newcomb's situation, we maintain, depends on a comparison of what would happen if one took both boxes with what would happen if one took only the opaque box. What the agent knows for sure is this: if he took both boxes, he would get a thousand dollars more than he would if he took only the opaque box. That on our view makes it rational for someone who wants as much [money][3] as he can get to take both boxes, and irrational to take only one box (Gibbard and Harper 1978, 155).

While this reasoning seems compelling, it nonetheless flies in the face of the stipulated facts; the vast majority (99%) of those who choose one box receive $1,000,000 while those who choose two boxes do not. As a matter of stipulated fact, those who choose one box receive more money, the agreed upon objective in the Newcomb scenario. Why would the rational choice lead to the undesired outcome? We argue *pace* Gibbard and Harper that the rational choice is to choose one box, once rationality is properly understood in this context in terms of psychological coherence.

---

[2] This sentiment reflects the dominant strategy (see section 3 below) and is clearly expressed in (Joyce 1999, 153): "... the 'If you're so smart why ain't you rich?' defense does nothing to let [the one-boxer] off the hook; she made an irrational choice that cost her $1,000."

[3] The original text reads "...as much much as he can get..."

Gibbard and Harper's analysis of Newcomb's Problem, which has become the canon, depends essentially on their assumption that a participant's decision is causally independent of what is in the opaque box. Only on the assumption of independence does it follow that choosing two boxes is the dominant strategy, i.e. that "What the agent knows for sure is this: if he took both boxes, he would get a thousand dollars more than he would if he took only the opaque box" (ibid.). David Lewis shares this intuition. "We [two-boxers] are convinced by counterfactual conditionals: If I took only one box, I would be poorer by a thousand dollars than I will be after taking both" (Lewis 1981, 377). It is this assumption we deny. Before turning to our argument we address some preliminary assumptions.

4. About the Predictor

The Newcomb Problem stipulates that there is a highly accurate predictor of a player's choice who determines the contents of opaque box A based on that prediction. In assessing what choice is rational for a player in the Newcomb Problem, we do not address how one might rationally come to believe that there is such an accurate predictor.[4] For our analysis, we assume that a player can take the high success rate of the predictor for granted and, furthermore, assume that this stipulated success is not contingent to her actual predictive success[5], which then might be due to nothing more than lucky guesses. In such a case previous predictive successes are no guarantee of continued success and so should not be a factor in reasoning about likely outcomes and we agree with the standard analysis that the rational choice is to take two boxes. The interesting case is when the predictor's success is nomic, hence counterfactual supporting. The player need not know how the predictor is so successful only *that* she is almost always correct in her predictions and it is no accident that she is.

Assuming the predictor's success is not contingent reveals an overlooked way in which a player's rationality is relevant to an analysis of the Newcomb Problem. While we do not know how the predictor divines a player's choice—number of philosophy courses, brain scans, or how many children one has—there must be telltale signs as to what the player will ultimately choose to which the predictor is sensitive, even if the player herself is unaware of those signs and changes her mind several times before the final decision. That is, we take the stipulated facts about the predictor's success as evidence for a causal process leading to a player's decision that the predictor can foresee:[6] the very causal process upon which the player's rationality supervenes.[7]

---

[4] As Bar-Hillel and Margalit (1972) note, intuitions about the Newcomb Problem might be driven in part by a general skepticism about the possibility of such a reliable predictor. Interestingly, Nozick (1993) reanalyzes conflicting intuitions about what is rational in terms of a decision value that depends in part on confidence in the predictor.

[5] We also have nothing to say about a supernatural predictor.

[6] It is the causal process that is essential to our analysis. The probability of successful prediction indicates how likely the predictor is to foresee that process, so for our analysis the exact value (99%) does not matter.

[7] So what of the player's free will? We take the stipulated facts to be inconsistent with at least certain libertarian notions of free will, in particular, the complete independence of a decision from any prior event. On such accounts of free will the Newcomb Problem is incoherent (or requires a supernatural

5. *Rationality and Coherence*

Psychological coherence is a hallmark of rationality; indeed, this is something close to a tautology. We speak both of people and behaviour as being rational, but in order of explanation the rationality of a person is primary. A person is rational when her intentional states are mostly consistent with each other and with evidence she encounters. Rationality can tolerate some local inconsistencies because we are not cognitively closed (i.e. aware of all of the entailments of our intentional states) but rational people typically update their beliefs in some way so as to eliminate or at least circumscribe contradictions they become aware of. On the other hand, evidently inconsistent beliefs or systematic unresponsiveness to evidence are characteristic of irrationality and are often assessed as psychological disorders such as schizophrenia, dementia, OCD, etc.[8], which in the extreme preclude intentional characterization altogether. The broad consistency of a rational person's intentional states provides the framework by which she can act for reasons, minimally to satisfy her desires given her beliefs. Thus, for an action or decision to be rational, reasons for it must cohere with a rational person's set of intentional states. If it fails to so cohere it is irrational for that agent.

Already our analysis reveals that, strictly speaking, a decision in isolation is neither rational nor irrational. Nonetheless, in many instances the background of intentional states with which a decision must cohere can be left implicit, and in those cases we can speak as if an isolated decision is rational or irrational without confusion. Most scenarios considered in decision theory are like this. After all, the background states are such mundane things as believing that the words in the language of communication have their standard meanings, that $1,000,000 is more money than $1,000, and desiring to get the biggest payout possible. However, we argue that Newcomb's Problem is not such a case; ignoring psychological coherence essential for rationality leads to the paradoxical results of standard analyses.

A consequence of psychological coherence is that some intentional states can be reliably predictive of other intentional states and subsequent decisions, without impugning an individual's autonomy (see footnote 7). For example, a person's political views tend to cluster as either liberal or conservative; someone with leftist views is likely to support gun registration legislation; someone on the right is likely to oppose state funded abortion. A person's background psychological states, including her beliefs and desires, short and long-term memory, reasoning abilities, etc. determine not only her space of rational decisions but also her dispositions to choose among those possibilities, often making reliable prediction possible even without supposing the fantastic divination powers of the oracle in the Newcomb Problem.

---

predictor). We note the similarities between an extremely reliable predictor impugning free choice and scholastic debates about the consistency of human free will with God's omniscience. (Thanks to Tom Lennon for pointing this out.) The Newcomb Problem's paradox concerning rational decisions is not, however, the same as the traditional theological paradox of how to reconcile free will with God's omniscience.

While issues of free will are clearly relevant to decision theory, further discussion is beyond the scope of this paper.

[8] We are not suggesting that irrationality is either necessary or sufficient for having a psychological disorder, only that they are often correlated in order to highlight the importance of psychological coherence for rationality. (Bortolotti 2013) discusses these relations.

Of course, in the Newcomb Problem we are unaware of how a player's standing psychological states connect with her final decision, so we can't see how those states cohere with any particular decision, as we can with say a cluster of political beliefs. But the predictor can. To make sense of the stipulations, there must be a causal chain from a player's state at the time of her interview with Cassandra to her final decision. Furthermore, however that chain is characterized, for her reasoning process to be relevant in determining her choice, which it must for the decision to be rational, it must supervene on at least some parts of that causal chain such that the supervening psychological states cohere.

6. Our Argument

In order for a participant's choice to be rational, she must be rational; her intentional states must cohere—again the alternative is that they do not cohere, i.e. they do not hang together by reason. How a person is inclined to choose, given her psychological profile, determines very reliably both what she will, in fact, do and what is in the opaque box. The coherence constraint on making a rational decision ensures that her choice is not independent of her psychological make-up, and her psychological make-up also influences the predictor, thereby linking her choice to what is in the opaque box. Once it is clear that a player's choice and what is in the box are dependent, the rational choice is to choose one box, since that is how the $1M gets in the box. The predictor mediates between the player's internal states and the external situation.

7. A Useful Analogy

Consider an analogy. Golf instructors emphasize the importance of the follow-through swing to hit a golf ball correctly. However, watching professional golfers in slow motion, the ball is in flight before the follow-through portion of the swing occurs. How then, short of some strange backwards causation, can the follow-through influence the flight of the ball? The answer, of course, is that the initial portion of the swing is not independent of the follow-through; components of a swing cannot be performed in isolation from each other, so unless the swing is such as to end with a good follow-through, the club will not strike the ball properly. Human physiology may just be such that without a proper follow-through it is not possible to hit a good golf shot. Similarly in the Newcomb Problem, given the stipulated conditions under which the predictor places a million dollars in the opaque box and the coherence constraint on a decision being rational, the only reliable way to get the million dollars in play is to be disposed to choose only the opaque box *and to follow through on that commitment*. The entire process from interviewing with the oracle/predictor Cassandra to choosing one or two boxes is a unit, like a golf swing, held together by a player's rationality. And like the golf swing, the follow-through is an essential part of the process, for without it the player will not strike the predictor in the right way for her to place the $1,000,000 in the opaque box.

Now it might be objected that we are not appreciating the force of the fact that in the Newcomb scenario, taking both boxes is a *dominant strategy*. That is, no matter what the current state of the world is (i.e., whichever way Cassandra chose earlier), the expected

utility of two-boxing is higher than that of one-boxing. Setting utility equal to dollars, for simplicity, the utility of two-boxing if Cassandra put the money in the box is $1,001,000, *vs* $1,000,000 for one-boxing; and if she did not, it is $1000 *vs* $0. So no matter what we take the probability of her having placed the money in the box to be, the expected utility of two-boxing is higher than that of one-boxing, and by precisely $1,000. So, clearly, the only rational decision is to take both boxes.

To us, what this argument demonstrates is simply that a treatment of the problem that considers the in-game choice to be independent of the earlier interview leads to contradictory results. Again the golf swing analogy is illuminating. Imagine a golfer who has over-stretched a shoulder muscle and is afraid of further damaging it—a likely outcome if she swings with a robust follow-through. *Ceteris paribus* what she would like to do is swing perfectly until the ball has left the club head, and then immediately stop putting effort into the swing, coming to a gentle halt with no over-extension. So our injured golfer should, it seems, stop putting any effort into her swing as soon as it passes the place where the ball lay (and hence has already hit and sent the ball on its way). It is the dominant strategy because the world could be one of two ways: If the ball has already been hit well, then there is no downside to abandoning the follow-through, and this has the advantage of avoiding risk of further injury. On the other hand if the ball has not been struck well at that point, a vigorous follow-through is not going to help—the ball has left the club, after all! So the only rational thing to do is swing hard at first, and then abandon the follow-through come what may. It should be clear what will happen to this golfer: she'll hit bad shot after bad shot, and go home scratching her head about where decision theory led her astray. The problem, again, was in taking the follow-through to be independent of the first half of the swing: because it is not, one has to view the whole swing as one decision-act in order for the theory to not lead one astray.

8. Further Considerations: Backtracking

We admit that our analysis will seem counterintuitive to some, so let's look at it more closely. We are denying that if a participant who chose one box and received $1,000,000 had chosen two boxes she would have received $1,001,000 and that a player who chose two boxes and received $1,000 would have received nothing if she had chosen only one box; and the basis of our denial is her rationality.[9] First consider a slightly modified case in which the predictor is perfect. Gibbard and Harper claim "The argument that the U-utility of taking both boxes exceeds that of taking only one box goes though unchanged" (1978, 154).[10] Intuitions divide here;[11] the U-utility calculation depends on the above counterfactual—namely, "What the agent knows for sure is this: if he took both boxes, he would get a thousand dollars more than he would if he took only the opaque box" (ibid)—which

[9] Our argument is not merely evidence based; the stipulated facts indicate to us that there is a causal connection between the decision and what is in the box because she is rational.

[10] U- utility is Gibbard and Harper's calculation of expected outcomes in which two-boxing always dominates one-boxing.

[11] (Ahmed 2015) discusses and rejects a discontinuity if prediction is perfect, that being the only case in which one-boxing is rational.

in turn depends on the independence of what is in the box from the player's decision. In the case in which the predictor is perfect, this assumption is manifestly unjustified. Under no circumstances *can* a participant receive $1,001,000. She either picks one box and wins $1,000,000 or two boxes and receives $1,000: guaranteed! So conceiving of the decision situation as one in which a million dollars either is or is not in the opaque box already, making it rational to choose both boxes for maximum payout, is to misconceive the situation, highlighting the fact that the decision and what is in the opaque box are not independent. What is true counterfactually in the case of a stipulated perfect predictor is that had a person who received a million dollars by choosing one box chosen two boxes, the opaque box would have been empty and she would have received $1,000. Likewise a person who gets $1,000 by choosing two boxes would have gotten $1,000,000, not $0, had she chosen just one box. The actual Newcomb case deviates from this case only slightly, when the predictor makes an error; so in the actual scenario too, apart from rare error cases, counterfactually if a person who chose one box and received $1,000,000 were to have chosen two boxes, she would have gotten $1,000.[12] There simply is no reliable means to obtain $1,001,000. The maximum payout that can be reliably obtained is $1,000,000, so the rational strategy is to make $1,000,000 the goal and do what is required to obtain it.

Some diagnosis of the diverging intuitions may be helpful here. Standard analyses consider whether the million dollars is in the opaque box or not, in isolation, and ask in that situation what will get the most money, with no consideration for how a participant influences the situation she is in. "We two-boxers think that whether the million already awaits us or not, we have no choice between taking it and leaving it" (Lewis 1981, 377). As advocates of the dominant strategy analyze the problem, the predictor is irrelevant.[13] Indeed, their analyses would apply equally if the participant fell ill just before airtime so her friend, who was not predicted, was sent in to play for her. In this substitute player scenario, the friend's decision really is independent of the process determining what is in the opaque box, so for her we agree it certainly is rational to take two boxes, hoping her friend was an unwavering one-boxer. But the actual participant is not in these circumstances.[14] Her rational psychological make-up determines both how she will be predicted and what she will choose. Her freedom to choose does not entail that her choice is independent of her past, and indeed it cannot be if she is rational.[15] If a participant's psychological make-up is such that she will not follow through in choosing only box A at the time of the decision, she will not strike Cassandra in the right way, who in turn will not place the $1,000,000 in the opaque box.

Another possible concern with our analysis is that we explicitly endorse what Lewis (1979) calls a backtracking counterfactual. We are taking it as true that had a person who chose one box and received $1,000,000 chosen two boxes she would have received only $1,000. Our reasoning is that because she is rational her choice and what's in the box are

[12] Our reasoning here is in agreement with Horgan's (1981).

[13] The case parallels the Monty Hall Problem in that ignoring how the circumstances arise leaves out essential information for calculating the correct expected utility.

[14] We see it as a virtue of our account that it clearly distinguishes what we take to be different scenarios that receive the same treatment on standard analyses.

[15] Fischer (2001) argues for the dominant strategy in response to Carlson (1998) on the grounds that our free choice is constrained by the actual past; however, his reasoning still requires the choice and prediction to be independent, which we deny for any rational agent. (See footnote 7).

not independent, so in order for her to counterfactually choose two boxes the causal antecedents would have to be different in ways that would also affect the predictor and hence what is in the opaque box. Lewis acknowledges that in special circumstances marked by their own peculiar grammar ("what would have had to have been the case such that...") backtracking counterfactuals are appropriate, but says we default back to standard contexts in which counterfactual dependence is asymmetric, the past being independent of the future. "Under this standard resolution, back-tracking arguments are mistaken: if the present were different the past would be the same, but the same past causes would fail somehow to cause the same present effects" (Lewis 1979, 457).

Lewis reasons that the nearest possible worlds are ones in which, from our point of view, small miracles, or violations of scientific laws, occur just before the critical moment in question, so that the past remains constant. "The deterministic laws of w_0 are violated at w_1 in some simple, localized, inconspicuous way. A tiny miracle takes place" (468). By contrast, backtracking counterfactuals of the sort we endorse require significantly more change, possibly reaching back to the origins of the universe. "But, under the backtracking resolution, the being's predictive correctness is a more important parameter of similarity than is maximization of the spatiotemporal region through which perfect match of particular fact prevails" (Horgan 1981, 336-7). Whatever the general account of counterfactuals may be, Lewis's account does not correctly capture the situation in the Newcomb Problem. Recall, the counterfactual in question is that were someone who took one box and won a million dollars to have taken two boxes, she would have won an extra thousand dollars. So the tiny miracle that Lewis describes, taken in the context of the Newcomb Problem, would change a committed one-boxer into a two-boxer at the last instant. Such a change would make her decision inconsistent with her own psychological history, which causes her actual choice, hence irrational. Furthermore, Lewis's analysis of the counterfactual leads to the following dilemma, anticipated by Horgan (1981). The small change either affects the predictor or it does not. If the change does not affect the predictor, she will make the same prediction as in the actual case, which will now be incorrect, contrary to the stipulated success rate of the predictor (see section 4 above). If, on the other hand, the prediction were to change along with the choice, the opaque box would be empty, as we claim. Either way, Lewis's account of the counterfactuals cannot be correct in the Newcomb scenario. "I think that Evidentialists typically concede too much to their Causalist opponents, in granting them the counterfactuals on which the charge that one-boxing is irrational always depends" (Price 2012, 507).

Several authors (Ahmed 2014, Fischer 2001, Vinci 1988) argue that backtracking counterfactuals are not relevant for deciding the Newcomb Problem, though they do not agree about the resolution. Others (Cantwell 2013, Hunter and Richter 1978) highlight the importance of how the problem is represented and the context sensitivity in interpreting counterfactuals. Indeed, from a semantic point of view, standard (fixed past) and backtracking counterfactuals are both legitimate uses of counterfactual conditionals. However, as Dummett (1993) points out, the Lewis-style interpretation is appropriate use of language after a decision is made and not a guide as to what it is rational to do. (Note, Dummett stipulates the possible contents of the opaque box as $10,000 rather than $1,000,000):

> *After* I have done it [chosen one box], the rules governing the assertion of counterfactual conditionals may entitle me to assert, "If I had taken both boxes I should have got $11,000"; but that

> is only a remark about our use of counterfactual conditionals. *Before* I make my choice I should be a fool to disregard the high probability of the statement, "If I take both boxes, I shall get only $1,000". That is not merely a remark about our use of the word "probability", nor even about our use of the word "rational", but about what it is rational to do (Dummett 1993, 375, emphasis in original).

9. Spohn's Analysis: Common Cause

As we analyze the problem, a person's prior psychological state is a common cause of both what is in the opaque box and how she will choose, linking them. Interestingly, Nozick, in his concluding footnote, recognizes that such a common cause may play a role in explaining Newcomb's Problem.

> But it also seems relevant that in Newcomb's example not only is the action referred to in the explanation of which state obtains (though in a nonextensional belief context), but also there is another explanatory tie between the action and the state; namely, that both the state's obtaining, and your actually performing the action are both partly explained in terms of some third thing (your being in a certain initial state earlier). A fuller investigation would have to pursue yet more complicated examples which incorporated this (Nozick 1969, 146, note 22).

Wolfgang Spohn (2012) offers a defense of one-boxing in which the common cause is the decision itself, which takes place at the time of the interview. We are broadly sympathetic to his analysis and agree with the strategy he prescribes. He does express some reservations at artificially placing the decision at the time of the interview. "Being committed or decided all along without ever having reflected on the matter? This sounds strange, and this may be the weak part of my account of NP, but, as I would insist then, the only weak part" (Spohn 2012, 103). To relieve this tension we suggest that it is more appropriate to speak of a decision process rather than a decision as a momentary event. As Spohn recognizes, what holds this complex causal nexus together is the participant's rationality. "You are decided early enough to one-box, simply by being rational, and this influences the predictor's prediction, presumably simply by his observation of your consistent and continuous rationality" (ibid.). His reason for putting the decision at the time of the interview is to block a screening off effect of any earlier common cause by a participant's self-awareness of her intention to one or two-box.[16] However, the consistency and continuity of a rational decision process do not admit such a break. The process is a unit, like a golf swing, so placement of *the* decision at any point within the decision process is an artificial stipulation.[17, 18] Nothing rules

[16] See (Eells 1982) for detailed discussion. For detailed discussion about when one can hold credences about what one will do see (Hájek 2016).

[17] Similarities between our reasoning here and Dennett's account of consciousness are not accidental.

[18] Our analysis may seem too restrictive to accommodate variations of the problem, such as when the prediction is made in the remote past. Our suspicion is that such scenarios are designed to press the intuition that the decision and prediction, hence what is in box A, are independent. However, we maintain that even when the prediction is in the remote past, the predictor is sensitive to a causal chain that culminates in the final decision, though, of course, only the smallest part at the very end of the chain will correspond to intentional states in the player's decision process. The common cause occurs in the

out that a participant might reflect about what to do after the interview, possibly changing her mind several times. Indeed, this is to be expected in problem scenarios that generate paradoxes through reflective self-reference, such as the Newcomb Problem. Thus a friendly amendment to Spohn's formalism, suggested by a correct understanding of rationality and its role of binding events into a decision process, is to re-interpret his common cause as the first stage of the decision process culminating in the act of taking one or two boxes. This amendment removes any artificiality in Spohn's analysis and is consistent with his own rationale for placing the decision at the beginning of the process. "...your introspection will reveal that often your reflection does not issue in a decision, but rather finds that you were already decided or committed" (ibid.). With this amendment our account and Spohn's are in agreement.

10. Clarifications

Note that we are not defending an evidence-based decision theory *per se*. Rather, the evidence indicates that there is a causal relation to be explained and puts constraints on what that explanation might be. As a result, the correct expected utility calculation in the Newcomb situation is more complex than what the objector imagines; it involves a decision process that takes place over (at least) the interval from when the game first begins until the final choice is reported, in which seemingly independent events are linked by a common cause. The causal decision theory (CDT) analysis goes wrong in treating the process as a single event occurring only at the last moment.[19] Of course, the player will find herself in the moment of truth. The opportunity to influence the predictor is gone and she has to decide what to do at that very moment, not lament about what she wishes she had done. The ideal course of action for playing the game is to somehow convince Cassandra that you will choose one box but actually choose two. The trouble with this course of action is that although we do not know her means, Cassandra is rarely fooled. And as the Newcomb Problem is standardly set up, this success is not a contingent fact about strategies that participants happen to have used. Even convincing yourself that you will be a one-boxer won't be enough. Consider the second-guessing-one-boxer, who convinces herself that she will choose one box. Having done so, at the moment she chooses she reasons that since she herself was convinced she would choose one box, Cassandra must have been similarly convinced and put the $1,000,000 in box A. But as with all other strategies this will have telltale signs that the participant will choose two boxes in

remote past in that, foreseeing the causal chain, the predictor links the contents of box A to the final choice. If this seems farfetched, note that the scenario already requires supposing a predictor in the remote past who can predict that such a game will be played at a particular time and location in the distant future, based only on her observations in the remote past. In such a case supposing she can also predict the choice to be played does not seem so fanciful after all.

[19] It seems to be the reflexive nature of the decision, which depends on reasoning about the prediction of that very decision that generates the paradox. In such cases, we prescribe Spohn's (2012) formalism, interpreted as a decision process, to model the situation. Other cases, especially those involving common causes, may also require treating decisions as processes, but a general analysis of when to do so is beyond the scope of this paper.

the end, so even if everyone adopted this second-guessing-one-boxer strategy, Cassandra still would be 99% accurate. A last minute change of heart will be predictable and box A will be empty. The upshot is that in the final instant, a player must realize that whatever she does next will follow from a process that left telltale signs the predictor used to determine what is in the box, making it rational to choose only one box. The player (and we) does not know herself how her psychological states cohere with her decision, only that they do. But the predictor knows us better than we know ourselves. A player making the rational choice exploits that fact to make it likely[20] the predictor put the $1M in box A; the $1,000 is a seductive trap that must be avoided.

Suppose the game were varied so that players could simply take $1,000,000 from box A and go home or first take $1,000 from box B and then open box A, with the catch that box B's lid was hooked up (in advance and preset) to a mechanism that 99% of the time released a trap door so the million dollars in box A would drop away and be lost. We suspect that very few two-boxers would be tempted by the $1,000 in this trap-door variation, even if the mechanism was preset based on a reliable prediction of what the player would choose; yet the actual Newcomb situation is much closer to this scenario than the way two-boxers analyze it, as a choice of one or two boxes whose content is already set. The game begins at the interview and proceeds through a complicated causal decision nexus captured in Spohn's (2012) common cause model. Two-boxers lament that they "were never given any choice about whether to have a million" (Lewis 1981, 377). They find fault with the game for rewarding irrationality. "We take the moral of the paradox to be something else: If someone is very good at predicting behavior and rewards predicted irrationality richly; then irrationality will be richly rewarded" (Gibbard and Harper 1978, 153). Our response is roughly that in a situation in which CDT-irrationality is richly rewarded the rational thing to do is act CDT-irrationally. "Any policy compliance with which has a strong possibility of not yielding the greatest advantage simply cannot be that which a rational agent will adopt" (Dummett 1993, 375). Less roughly, given the goal of getting as much money as possible, one should do whatever will reliably get you the most money; and that is the rational course of action, even in decision theoretic terms. Thus, in the Newcomb Problem the rational action is committing to taking one box at the outset and following through on that commitment. A player's final choice does not cause there to be a million dollars in the box nor make her the type of person who (earlier) strikes the predictor as a one-boxer or a two-boxer. It reveals her nature, possibly even to herself, though not, of course, to the predictor.[21] Can she resist low-hanging fruit? The game rewards disciplined rationality.

[20] Essentially, knowing there is a common cause of what is in box A and her decision, a rational player picks one box and trusts the predictor got it right. The probability of the predictor's success is relevant insofar as it is the basis of that trust.

[21] Nozick makes an even stronger claim. "Although performing an action of the sort that would be done by a certain kind of person may not *cause* the agent to be this kind of person, it may symbolize his being that way, be some evidence that he is, and have the causal consequence of making it easier for him to maintain an *image* of himself as being of that kind. This last is a real causal consequence of an action and may have significant utility" (Nozick 1993, 49, emphasis in original).

11. The Look-First Scenario

We conclude with one final variation that Gibbard and Harper presented with the intention of highlighting that the original Newcomb game rewards irrationality. In this new look-first scenario, after the same prediction process one is allowed to look in the opaque box and then decide whether to take the $1,000. Since, on a sequential decision-theoretic analysis there is no apparent cost for the one-boxer after the fact to take the extra thousand dollars, it seems the rational strategy is to take the extra $1,000 after finding $1,000,000. Again the conclusion is compelling if one mistakenly assumes that taking the $1,000 after the fact is independent of earlier events; but in fact the scenario is inconsistent with the stipulated parameters of the Newcomb situation, highlighting the dependence of the two choices.

The two problems are importantly different. The look-first scenario admits four possible strategies, as opposed to the two strategies of one-boxing or two-boxing in the original Newcomb Problem. In addition to the original two strategies, players might choose a strategy of reacting to what the predictor does. A player might choose to be a cooperative responder, settling for the $1,000,000 if it is there, as a one-boxer would, but taking the $1,000 consolation if the million is not in box A, as a two-boxer would. Interestingly, whether the million dollars is in box A or not, the predictor cannot be wrong about the cooperative responder, and so Cassandra gets to choose what this player will win. Assuming her choice is random, half of the time cooperative responders win $1,000,000, the other half they settle for $1,000. The final strategy is the problematic one for this scenario. A player can choose to be an uncooperative responder, or a defier. The defier does the opposite of how she is predicted to behave. If the million dollars is in box A, a prediction of one-boxing, the defier takes both boxes getting $1,001,000. If box A is empty, a prediction of two-boxing, the defier, perhaps out of spite, takes nothing. So the predictor can never be right about a defier, contrary to the stipulation that she must be correct 99% of the time. Rather than showing that in the original Newcomb Problem two-boxing is the more rational strategy, the incoherence of this scenario makes manifest how crucial the stipulated success rate of Cassandra is for the correct analysis of the original Newcomb Problem—precisely what the defenders of two-boxing set aside as irrelevant.[22]

12. Conclusion

We have used Newcomb's Problem to highlight an often-overlooked feature of rational decisions, namely that their rationality consists in cohering with a background of coher-

[22] If further stipulations are added about the number of defiers to keep the predictors overall success rate very high and the predictor chooses at random in both responder scenarios since she is either guaranteed to be correct or incorrect, one-boxing is still the best strategy in the look-first scenario since it makes it most likely the $1,000,000 will be in play. Perhaps no one is so "hyper-rational" as to actually one-box in the look-first scenario, but overwhelmingly the alternative is to find box A empty. "When those who leave the thousand dollars are asked later why they do so, they say things like 'If I were the sort of person who would take the thousand dollars in that situation, I wouldn't be a millionaire'" (Gibbard and Harper 1978, 154). Just so, Socrates.

ent psychological states. Decision-acts are like golf swings: the right way to think of them is as a unit, where the earlier parts of the action are simply not independent of the later parts, nor is the final upshot of the action (i.e. success or failure). A golfer *can* suddenly stop swinging hard an instant after the club strikes the ball, but she is fooling herself if she thinks that, in a case where she does that, the first half of the swing can be exactly the same as it would have been in a case where she intended all along to have a good follow-through *and did so*. Applied to the Newcomb Problem, this makes the rational strategy to choose one box. Analyses that determine two-boxing to be a dominant strategy ignore the role of the predictor that establishes non-mysterious causal connections via a common cause linking the contents of the opaque box to what a player decides.[23] "I do recommend acting *as if* one's present choice could causally influence the being's prior prediction, but my argument does not presuppose backward causation" (Horgan 1981, 340-341, emphasis in original). If what is in the opaque box were independent of the player's choice then the best strategy would be to two-box; however, the player's rationality guarantees they are not independent.

Acknowledgements

We would like to thank Timothy Kenyon and Dustin Locke for comments on earlier drafts of this paper and especially William Harper for numerous discussions of these issues and comments on the penultimate draft.

REFERENCES

Ahmed, Arif. 2014. Causal decision theory and the fixity of the past. *British Journal for the Philosophy of Science* 65/4: 665-685.

Ahmed, Arif. 2015. Infallibility in the Newcomb problem. *Erkenntnis* 80/2: 261-273.

Bar-Hillel, Maya and Margalit, Avishai. 1972. Newcomb's paradox revisited. *British Journal for the Philosophy of Science* 23/4: 295-304.

Bortolotti, Lisa. 2013. Rationality and sanity: The role of rationality judgments in understanding psychiatric disorders. In K.W.M. Fulford, M. Davies, R.G.T. Gipps, G. Graham, J.Z. Sadler, G. Stanghellini, and T. Thornton, eds., *The Oxford handbook of philosophy and psychiatry*, 480-496. Oxford: Oxford University Press.

Cantwell, John. 2013. Conditionals in causal decision theory. *Synthese* 190/4: 661-679.

Carlson, Erik. 1998. Fischer on backtracking and Newcomb's problem. *Analysis* 58/3: 229-231.

Dummett, Michael. 1993. Causal loops. In *The seas of language*, 349-375. Oxford: Oxford University Press.

Eells, Ellery. 1982. *Rational decision and causality*. Cambridge: Cambridge University Press.

Fischer, John Martin. 2001. Newcomb's problem: a reply to Carlson. *Analysis* 61/3: 229-236.

Gibbard, Alan & Harper, William. 1978. Counterfactuals and two kinds of expected utility. In A. Hooker, J. J. Leach & E. F. McClennen , eds., *Foundations and applications of decision theory*, 125-162. D. Reidel.

Hájek, Alan. 2016. Deliberation welcomes prediction. *Episteme* 13/4: 507-528.

[23] In advocating the case for a common cause we are meeting McKay's (2004) prescription without cheating or backwards causation.

Horgan, Terence. 1981. Counterfactuals and Newcomb's problem. *Journal of Philosophy*, 78/6: 331-356.
Horwich, Paul. 1985. Decision theory in light of Newcomb's problem. *Philosophy of Science* 52/3: 431-450.
Hunter, Daniel and Richter, Reed. 1978. Counterfactuals and Newcomb's paradox. *Synthese* 39/2: 249-261.
Joyce, James. 1999. *Foundations of Causal Decision Theory*. Cambridge University Press.
Lewis, David. 1979. Counterfactual dependence and time's arrow. *Nous* 13/4: 455-476.
Lewis, David. 1981. Why ain'cha rich? *Nous* 15/3: 377-380.
McKay, Phyllis. 2004. Newcomb's problem: the causalists get rich. *Analysis* 64/2: 187-189.
Nozick, Robert. 1969. Newcomb's problem and two principles of choice. In N. Rescher, ed., *Essays in honor of Carl Hempel*, 114-146. D. Reidel.
Nozick, Robert. 1993. *The nature of rationality*. Princeton University Press.
Price, Huw. 1986. Against causal decision theory. *Synthese* 67/2: 195-212.
Price, Huw. 1991. Agency and probabilistic causality. *British Journal for the Philosophy of Science* 42/2: 157-176.
Price, Huw. 2012. Causation, chance, and the rational significance of supernatural evidence. *The Philosophical Review* 121/4: 483-538.
Spohn, Wolfgang. 2012. Reversing 30 years of discussion: Why causal decision theorists should one-box. *Synthese* 187/1: 95-122.
Vinci, Thomas C. 1988. Objective chance, indicative conditionals and decision theory; or, how you can be smart, rich and keep on smoking. *Synthese* 75/1: 83-105.

Christopher Viger is a philosopher of mind, psychology and cognitive science. His work applies the insights of cognitive neuroscience to understanding the relation between language and thought.

Address: Department of Philosophy, Western University, 1151 Richmond Street. London, Ontario, Canada. N6A 3K7 Email: cviger@uwo.ca

Carl Hoefer is a philosopher of science whose research areas include determinism, causality, chance, laws of nature, and physical theories.

Address: ICREA, Pg. Lluís Companys 23 (08010 Barcelona), Spain. Departament de Filosofia. Universitat de Barcelona. Carrer Montalegre 6, 4.ª planta (08001 Barcelona), Spain. Email: carl.hoefer@ub.edu

Daniel Viger holds a masters degree in philosophy and is currently a law student at the University of Western Ontario.

Address: Faculty of Law, Western University, 1151 Richmond Street. London, Ontario, Canada. N6A 3K7. Email: dviger2@uwo.ca